Psicológica (2011), 32, 403-421.

Assessing Measurement Equivalence in Ordered-Categorical Data

Paula Elosua[*]

University of the Basque Country

Assessing measurement equivalence in the framework of the common factor linear models (CFL) is known as factorial invariance. This methodology is used to evaluate the equivalence among the parameters of a measurement model among different groups. However, when dichotomous, Likert, or ordered responses are used, one of the assumptions of the CFL is violated: the continuous nature of the observed variables. The common factor analysis of ordered-categorical data (CFO) has been described in several works, but none evaluate its power and Type I error rate in the evaluation of measurement equivalence (ME). In this simulation study, we evaluated ME under four different conditions: size of group (300, 500 and 1000), type of DIF (thresholds, loadings), amount of DIF (0.25, 0.40), and equality/impact of the distributions. The parameters used for the data generation came from one scale with nine items with three ordered categories. The results were evaluated according to three decision rules: a) the significance of the difference in chi-square values obtained in two nested models, b) the significance of the difference in chi-square values between two nested models with Bonferroni corrections, and c) the difference between the values of the Comparative Fix Index (CFI) obtained in two nested models. The results showed good power as well as good control of the false positives for both the chi-square Bonferroni correction and CFI difference index.

The common factor linear model has been used with large success for modelling measurement and structural models in the framework of psychological and educational measurement (Bollen, 1989, Jöreskog and Sörbom, 1993). The model is extensively applied even when the basic assumption regarding the continuous nature of the indicator variables is violated. Most tests or questionnaires assessing psychological and educational latent variables use ordered-categorical response variables, such as dichotomous response items, Likert type items, or partial-credit items.

[*] This research was supported through grants (PSI2008-00856) by Ministry of Science and Innovation and by University of Basque Country (GIU09/22). Correspondence: Paula Elosua. University of the Basque Country. Avda. Tolosa, 70. 20018 San Sebastian, Spain. E-mail: Paula.elosua@ehu.es

Although the nature of the responses in these cases is discrete and there are important studies about models for ordinal variables when they are used as latent variable indicators (Bock and Aitkin, 1981; Christofferson, 1975; Jöreskog, 1990; Mislevy, 1986; Muthén, 1984), the approach most often used to model the ordered data is designed for continuous data (Muthén and Kaplan, 1985). From a theoretical point of view this is a methodological issue related to the application of incorrect models.

This issue is extended to the evaluation of factorial invariance. Assessing factorial invariance implies analyzing the mean and covariance structures of the data (MACS; Sörbom, 1974) across groups. The definition of factorial invariance involves the equivalence between conditional probabilities for observed outcomes given latent variable scores (Mellenbergh, 1989; Meredith, 1993) among groups. This statement is independent of the nature of the observed scores:

$$P\left(Y|\eta, g_1\right) = P\left(Y|\eta, g_2\right) \tag{1}$$

To assess measurement invariance, we need to check the parameters that determine the probability of observed scores among groups, in other words, the equivalence among the parameters of the measurement models among the groups (Meredith, 1964; 1993). It is a progressive study of covariance structures, where the same measurement model is estimated in two groups using multigroup confirmatory factor analysis. The invariance level reached depends on the invariance of the parameters used in the definition of the measurement model. In the framework of linear models, to guarantee that the same construct is assessed in different groups and that this construct has the same metrics characteristics among groups, we must evaluate the equivalence among the regressor coefficients (Λ), the intercepts (ν), and the equivalence among the residual variance/covariance matrix (Θ). The parameters that remain equivalent will determine the kind of invariance: configural, metric, strong, or strict (see table 1). The first, configural equivalence, is the simplest model of invariance. It assumes the equivalence between the basic configurations of the measurement model in both groups; that is, the factors are defined through the same items among groups. Metric equivalence adds a restriction to the previous model: the equality between the regressor coefficients or loading in both samples ($\Lambda_1=\Lambda_2$). The third one, strong equivalence, demands the equality among intercept parameters ($\nu_1=\nu_2$), and the highest model of invariance constrained to be equal to the error variances ($\Theta_1=\Theta_2$). Table 1 formally shows the invariance models.

To evaluate measurement invariance, multiple group analysis is typically performed, as well as the computation of the chi-square difference test for nested models. The chi-square value and degrees of freedom for the less restrictive model are subtracted from the chi-square value and degrees of freedom for the more restrictive model. Depending on the significance of the difference, we are able to conclude the invariance or non-invariance of the model across groups.

But models for continuous and ordered data are different; therefore, testing the invariance of the measurement parameters between groups (Meredith, 1993) implies testing different parameters for each model. In case of continuous response data or multivariate normal data, the parameters of interest are regression intercepts, factor loadings, and residual variance (table 1), but if there is ordered categorical data, the set of parameters to be assessed is different because thresholds need to be compared across groups. Using continuous linear models with maximum likelihood estimation to analyze ordered-categorical data potentially has several disadvantages and can leads to erroneous conclusions because the models are formally different, and the information on the focus of invariance or the source of unacceptable fit remains obscure (Lubke and Muthén, 2004).

Despite of those differences among models several papers have analyzed the power and Type I error associated with the common factor linear models for continuous data in assessing measurement invariance with ordered data (Elosua and Wells, 2008; Meade and Lautenschlager, 2004a; Stark, Chernyshenko, and Drasgow, 2006). They basically compare the power associated with procedures designed specifically for ordered data derived from Item Response Theory, as the Likelihood Ratio Test (Thissen, Steinberg andWainer, 1988) with that obtained using the MACS approach with maximum likelihood estimation. The conclusions among results are not consistent. Meade and Lautenschlager (2004a) and Elosua and Wells (2008) found more power associated with IRT procedures, especially when the lack of invariance or differential item functioning (DIF) was generated in thresholds (uniform DIF); However Stark, Chernyshenko, and Drasgow (2006) concluded more power for the MACS procedure. One of the reasons for these results may be the model used in data generation. The first two studies generated data from the IRT logistic models, and the last study generated data from continuos linear model.

There are few studies have analyzed the power in detecting measurement invariance using the common factor linear model for categorical outcomes. French and Finch (2006) studied the power and Type I error associated to the evaluation of measurement invariance on loadings with dichotomous items in multidimensional data. Using chi-square difference tests as criteria for evaluating measurement invariance their findings showed very low power in detecting differences in loadings (in the best condition the power doesn't reach 20%). Finch and French (2007) evaluated factorial invariance in an unidimensional dichotomously scored test by generating DIF on discrimination parameter. Even when the magnitude of DIF was big (0.6) the power didn't reach the 60% test This lack of power related with loadings was also reported by using models for continuous outcomes (Meade and Lautenschlager, 2004a). However there is not works analyzing the power of models for ordered data with unidimensional and polytomous data and it is clear we need studies about this issue (Millsap, 2005). Studies of this kind are important in order to evaluate the effectiveness, applicability and possible deficiencies of the model for categorical outcomes.

Table 1. Invariance models for continuous data

Invariance models		
$\sum_1=\Lambda_1\Psi_1\Lambda_1^T+\Theta_1$	$\mu_1=\nu_1$	Configural invariance
$\sum_2=\Lambda_2\Psi_2\Lambda_2^T+\Theta_2$	$\mu_2=\nu_2$	
$\sum_1=\Lambda\Psi_1\Lambda^T+\Theta_1$	$\mu_1=\nu_1$	Metric invariance
$\sum_2=\Lambda\Psi_2\Lambda^T+\Theta_2$	$\mu_2=\nu_2$	
$\sum_1=\Lambda\Psi_1\Lambda^T+\Theta_1$	$\mu_1=\nu$	Strong invariance
$\sum_2=\Lambda\Psi_2\Lambda^T+\Theta_2$	$\mu_2=\nu+\Lambda\alpha_2$	
$\sum_1=\Lambda\Psi_1\Lambda^T+\Theta$	$\mu_1=\nu$	Strict invariante
$\sum_2=\Lambda\Psi_2\Lambda^T+\Theta$	$\mu_2=\nu+\Lambda\alpha_2$	

With no more studies designed to assess the power and Type I errors associated to models for ordered outcomes we wanted to study this issue using a simulation that breaks some new ground in the field of measurement invariance. Our works extended French and Finch's (2006) research in several aspects; a) we simulated data for ordinal items with three response categories; b) we simulated non invariance in loadings and thresholds; c) we studied the effect of impact or different latent means for groups on the detection of measurement invariance, and finally d) we compared three criteria for evaluating measurement invariance; CFI difference tests, chi-square difference tests, and the Bonferroni correction.

To achieve this goal we started with a description of the model for ordered outcomes and the description of how the factorial invariance can be assessed.

Ordered-categorical models.

The common factor linear model for categorical outcomes is one extension of the common factor linear model for continuous data. Basically, it is assumed that the observed score (Y_{ij}) is determined by unobserved scores on the latent response variables (Y^*_{ij}). Those latent variables are continuous, and so, the observed measures can be viewed as discretized versions of the latent responses variables. This discretization depends on the latent threshold parameters (τ).

Let Y_{ij} be the score on the *j*th categorical measures for the *ith* person, and let *c* the number of response categories:

(2) $$Y_{ij} = m \text{ if } \tau_{m-1} \leq Y^*_{ij} < \tau_m$$

Where m = 1, … C, and [$\tau_{j0,}$ τ_{j1}, …τ_{jC}] are latent thresholds parameters for the *j*th variables. The thresholds partition the range y^* into C categories. The thresholds are therefore ordered, τ_0 $<\tau_1$ $<\tau_C$, where $\tau_0 = -\infty$ and $\tau_C = +\infty$, and they may differ across variables.

Given that the observed score Y_{ij} is assumed to be determined by unobservable latent continuous variable Y^*_{ij}. (we consider by simplicity the unidimensional model), we can write

(3) $$Y^*_{ij} = \nu_j + \lambda j \eta_{ij} + \varepsilon_{ij}$$

Where ν is the intercept of the regression, λ is the factor loading or regression coefficient, η_i is the factor variable, and ε_i is the residual term. This equation is equivalent in the continuous linear model (Jöreskog, 1971)

It is possible to represent the arithmetical means and variance of the continuous latent response variable by means of structural parameters.

(4) $$\mu^* = \nu + \lambda\alpha$$
$$\sigma^* = \lambda^2\psi + \theta$$

For these equations, α is the mean of η, ψ is the variance of η, and θ is the variance of the ε. Under this model the observed variable (Y) is modeled using the measurement parameters regarding the latent response variable (Y*), and the thresholds parameter (τ), which determine the values of the categorical response variable.

In this framework of common factor linear models for ordered categorical variables, the objective of this work was to investigate the power and Type I error rates for detecting a lack of measure equivalence (or differential item functioning in the framework of item response theory; Lord, 1980). In addition, we wanted to study the strategy for detecting DIF based on the likelihood ratio test. This is a common way to assess DIF in the framework of IRT (LR; Thissen, Steinberg, and Wainer, 1988), and also, we wanted to check different criteria for flagging DIF items. We used three different decision criteria: (a) the difference between two chi-square values tests belonging to nested models, (b) the Bonferroni correction, and (c) the difference in Bentler's comparative fix index (CFI; Bentler, 1990) between two nested models. According to this criteria, an item was flagged for DIF when the difference between two nested models in the comparative fix index was greater than 0.01 (Cheung and Rensvold, 2002).

METHOD

Detecting of DIF

The first step in the detection of DIF was testing the full invariance model. In this model all items parameters (thresholds and loadings) were constrained to be equal across groups[1]. This approach to assessing invariance is called the "constrained baseline approach", which is the opposite of the "free-baseline approach" (Stark, Chernyshenko, and Drasgow, 2006).

For each of the items in the test, a different model was evaluated. In each model the parameters of the evaluated item were freely estimated in the focal group. For identification purposes, the loading of the first item was fixed to one. Since this item is not evaluated, it becomes the referent item (Lubke and Muthén, 2004).

For each item and for each sample replication, the chi-square difference test (without correction and with Bonferroni correction) and the CFI difference index were computed between both the invariance model and the item-free model. We made eight comparisons (one for every item) for each data replication (we did not evaluate the reference item). This

[1] In this study, we followed the parameterization described by Lubke and Muthén (2004). We define $E(\eta)=\alpha=0$ so that $\mu=\nu$. Since y* is a latent variable, its metric is not determined and it is, therefore, common to standardize to $\nu=0$ between groups.

allowed us to analyze the power rates (correct detections) and false positives or Type I error rates. The power rate was defined by the number of times that the manipulated item was flagged across the replications. The Type I error rate was defined by the number of times that the non-DIF items were flagged across the replicas

An item was flagged for DIF if the difference in the chi-square values between two models was significant or the difference in the CFI index was greater than 0.01.

Study Design

The simulation study designed to assess the power and Type I error rate of linear models with categorical outcomes was made by defining and manipulating the following factors:

1 - Sample size:

Three sample sizes were selected for each group, reference group and focal group; (a) 300, (b) 500, and (c) 1000. Each was chosen respectively to represent a small, medium, and large sample size used in empirical research.

2 - Type of DIF:

Two types of DIF were used, both (a) DIF on thresholds and (b) DIF on loadings. The first condition involved uniform DIF and was defined by adding one constant to the thresholds of the DIF item in the focal group. Under equal loading, making focal group thresholds higher the item would be more difficult for focal group respondents. The second was the non-uniform DIF condition or DIF on loadings condition, where the factor loading of one item for the focal group was obtained by substracting a constant from the same item loading for the reference group.

3 - Amount of DIF:

Three levels of DIF were manipulated, (a) non-DIF (in these conditions the reference group and focal group loadings and thresholds parameters were set to be equal), (b) 0.25, and (c) 0.40. These values are the constant values that we added or subtracted to define the parameters for the focal group. We refer to these conditions as non dif, medium DIF, and large DIF.

The difference in non invariant item parameters was consistent with previous simulation works (Meade and Lautenschlager, 2004a; French and Finch, 2006).

4 - Amount of impact:

Two amounts of impact were examined, both (a) non-impact condition (the distribution of the reference group and the focal groups were set equal ($N_R(0, 1)$; $N_F(0, 1)$)) and (b) moderate impact (the mean of the focal groups was 0.5 standard deviations lower than the reference group ($N_R(0, 1)$, $N_F(-0.5, 1)$)).

5 - Decision rule to flag DIF:

Three rules were used to flag DIF items, (a) p=0.05 or Chi-square difference test without correction, (b) the Bonferroni correction, and (c) the CFI difference index.

In total, 30 different conditions were fixed to generate data under DIF conditions. Six non-DIF conditions were defined and three decision rules were evaluated. For each condition, 100 sample replications were generated.

Data

Realistic values for the generation of data were obtained using the estimated parameters of the one self-concept scale, AFA-A (Musitu, Garcia, and Gutierrez, 1997). The scale is a four-dimensional measure of the self-concept. For our purpose we used a unidimensional subscale composed of nine items with two thresholds. The use of scales with this number of items is common in psychological testing; we could cite for instance the different versions of the 16PF (Catell, 1989, 16PF-APQ; Schuerger, 2001). The response data were collected from 540 students (Elosua and López, 2008). The parameters used for the data generation can be found in Table 2. The fit of the data to the model was moderate (χ^2 =76.20; df=24; CFI=0.89; RMSEA=0.06).

In all conditions, the data were generated for two groups under a single-factor model with ordered outcomes. The model for the generation was the model implemented in Mplus3.11. The referent item for the model estimation was item 1, and the DIF item was number 9, the last item. Table 1 shows the parameters used in the data generation.

Analysis

The analysis of the data was carried out using Mplus 3.11 (Muthén and Muthén, 2004) using weighted least squares estimation (WLS). When ordinal data are analyzed weighted least squares (WLS) estimation (Muthén, du Toit and Spisic, 1997) is applied. That procedure uses polychoric correlations among items and is effective in estimating models with dichotomous and ordinal variables (Jöreskog, 1994).

Table 2. Item parameters for data generation

Item	Reference Group			Focal Group					
				DIF on λ		DIF on τ			
				Medium	High	Medium		High	
	λ	τ_1	τ_2	λ	λ	τ_1	τ_2	τ_1	τ_2
1	0.50	-1.2	1.1	0.50	0.50	-1.2	1.1	-1.2	1.1
2	0.21	-0.43	0.98	0.21	0.21	-0.43	0.98	-0.43	0.98
3	0.44	-0.61	1.2	0.44	0.44	-0.61	1.2	-0.61	1.2
4	0.50	-1.20	0.52	0.50	0.50	-1.20	0.52	-1.20	0.52
5	0.55	-0.76	1.02	0.55	0.55	-0.76	1.02	-0.76	1.02
6	0.70	-1.22	0.64	0.70	0.70	-1.22	0.64	-1.22	0.64
7	0.35	-0.31	1.2	0.35	0.35	-0.31	1.2	-0.31	1.2
8	0.37	-1.2	0.68	0.37	0.37	-1.2	0.68	-1.2	0.68
9	0.56	-1.2	0.80	**0.31**	**0.16**	**-0.95**	**1.05**	**-0.8**	**1.2**

Note: Numbers in bold are DIF items parameters.

For each replication nine models were analyzed. The invariance model or constrained baseline model, was used as well as eight more models, one for each item. The results of the analysis were extracted using R (Ihaka and Gentleman, 1996) , and the chi-square differences (without correction and Bonferroni corrected) and the CFI index difference were computed for each item in each replication in all evaluated conditions.

RESULTS

Tables 3, 4, and 5 show power and Type I error results for the analysis. Power represents the proportion of correctly flagged items across 100 replications in each condition. Type I error represents the proportion of times that no-DIF items were erroneously flagged. Note that this proportion was estimated across the replications and across all of the no-DIF items. We estimated the differences between nested models for each of the items belonging to the tests (DIF-item, rest of the items) and the baseline model (constrained model). The first item of the test was excluded from the analysis, which was fixed to one for identification purposes.

No-DIF condition

Table 3 presents the results for the no-DIF condition. The values in the table are the Type I error estimates for this condition. The column labeled with *p*=0.05 represents the values obtained by the difference between two chi-squares belonging to two nested models (invariance or constrained model versus “item free” model). These values were extremely big, ranging from 0.07 to 0.17. All were found to be greater than the nominal level. Even in the large sample condition, the reported value was 0.07. The Bonferroni correction reduced the percentage of false positives in all evaluated conditions. This correction eliminated Type I error, and all the values were below 0.03, even in the impact conditions. The CFI difference index showed the best control of the Type I error under all evaluated factors. Its value under the worst condition (small sample and impact) was 0.008. It is remarkable that neither the Bonferroni correction criteria nor the CFI difference index were influenced by the presence of distributional differences between groups.

Table 3. Type I error rates for no-DIF conditions

Impact	Sample Size	*p*=0.05	Bonferroni	CFI
None	300	0.16	0.03	0.004
	500	0.11	0.02	0.001
	1000	0.07	0.01	0.000
Impact	300	0.17	0.03	0.008
	500	0.12	0.02	0.004
	1000	0.07	0.01	0.000

DIF in thresholds

Power. (See table 4). The power of the evaluated three criteria was excellent in the high DIF condition (DIF 0.40). The rate was 1.00 for the chi-square difference in the presence or absence of impact. This rate was slightly reduced for Bonferroni correction and CFI differences in the condition of impact and small sample size, but still remained close to 1.00 (0.98 and 0.95, respectively).

When the amount of DIF was smaller, say 0.25, the performance of the evaluated criteria was a little different. The power increased as the sample size increased. Under the non-impact condition, the performance of the chi-square difference was close to 1.00, even in the small sample size condition (0.95). The rates are slightly reduced when the Bonferroni correction was applied. In this case the power for detecting DIF in small sample was 0.82, and increased when the sample size was medium (0.93). This pattern was followed also by the CFI difference index, but the power for this index was a little smaller. The values ranged from 0.64 in the small sample to 1.00 in the big sample condition, with a value of 0.74 for the medium sample condition.

When distributional differences were introduced in the analysis, the detection of power decreased. Under high DIF conditions all the indexes showed good behaviors. When the amount of DIF was 0.25 and the sample size was small, the power showed by CFI was only 0.62. This value increased to 0.72 in the medium size sample condition and raised to 0.77 when the sample was 1000. The power associated with Bonferroni corrections was not high when the size of group was small (0.76). In the medium sample size condition (N=500), the power was above 0.90 for the factor of a medium amount DIF and there was maximum power in the high DIF condition.

Type I error. (see table 4). The Type I error rates for the chi-square difference test was inflated in all conditions. It ranged from 0.14 to 0.28. These values increased as the amount of DIF increased. The use of the Bonferroni corrections, which is a stricter criterion for flagged items, reduced the Type I error. All the values reported were closed to 0.05, ranging from 0.04 to 0.09. They were not influenced by the presence of impact. The CFI difference showed values close to 0 in all conditions. This index was not affected for the amount of DIF. Even when the sample size was small (N=300), this index did not exceed the value of 0.02.

DIF in loadings

Power. The results of assessing non-uniform DIF are presented in Table 5. The three decision criteria correctly flagged the DIF items when the amount of DIF was 0.40. Only for the impact condition with a small sample size did the power of the CFI not reach the maximum value (this value was 0.94). When the amount of DIF was 0.25 the power rates increased as the sample size increased. For the large sample size, the chi-square difference tests (with correction and without correction) correctly flagged the DIF items. The power rate of the CFI difference index for this condition was 0.91. When the sample size was N=500, the rates related with the chi-square tests were good, bigger than 0.95 in all conditions (with correction, without correction, and under impact condition). These values were reduced when the sample size was 300. In this condition and the without impact condition, the difference without correction showed 0.98 power rate and the Bonferroni correction detected DIF 84 percent of the time. These values were smaller when distributional differences were present in the data. Under this adverse condition, the non-corrected chi-square power rate was 0.89, and the corrected rate was 0.78. The CFI difference index showed power rates smaller than the chi-square tests. The rates ranged from 0.78 to 0.91 for the equal ability distribution condition. These values decreased (0.64 to 0.89) in the presence of impact.

Table 4. Power and Type I error rates for DIF in thresholds

Impact	Sample Size	Amount DIF	Power			Type I Error		
			p=0.05	Bonferroni	CFI	*p*=0.05	Bonferroni	CFI
Non	300	Medium	0.95	0.82	0.64	0.20	0.07	0.02
		High	1.00	1.00	1.00	0.23	0.09	0.02
	500	Medium	0.99	0.93	0.74	0.17	0.04	0.003
		High	1.00	1.00	1.00	0.22	0.07	0.008
	1000	Medium	1.00	1.00	1.00	0.16	0.04	0.001
		High	1.00	1.00	1.00	0.28	0.06	0.001
Impact	300	Medium	0.95	0.76	0.62	0.22	0.05	0.01
		High	1.00	0.98	0.95	0.27	0.08	0.03
	500	Medium	0.99	0.90	0.72	0.16	0.03	0.005
		High	1.00	1.00	1.00	0.23	0.07	0.005
	1000	Medium	1.00	1.00	0.77	0.14	0.05	0.00
		High	1.00	1.00	1.00	0.23	0.08	0.00

The effect of the distributional difference between groups was bigger in the condition of small group (N=300) and medium DIF (0.25). In these cases, we reported values from 0.64 for the CFI index, 0.78 for the Bonferroni correction, and 0.89 for the no correction chi-square. This pattern was found also when the size of the groups was N=500. When the amount of DIF was big (0.40), the power was not influenced by the impact. The effect of the non equal distributions was bigger in the small sample conditions. The detection rule with less power was the CFI difference index. This index was close to 1.0 when the amount of DIF was big, but was reduced when the amount was 0.25.

Type I error. The Type I error control was not good for the chi-square difference test. The values ranged from 0.10 to 0.34. The inflated error was drastically reduced by the Bonferroni correction. For this index, when the sample size was medium (N=500) or high (N=1000), the values were below 0.05. When the sample size was small (N=300), those values were slightly bigger (0.08 and 0.11). Those values increased with the amount of DIF increased. The CFI difference index showed the best control of the Type I error. The obtained values were close to 0. Even in the worst condition (small sample, medium DIF, and impact) we reported a value of 0.06. We did not observe any influence of the impact over this index.

Table 5. Power and Type I error rates for DIF in loadings

			Power			Type I Error		
Impact	Sample Size	Amount DIF	*p*=0.05	Bonferroni	CFI	*p*=0.05	Bonferroni	CFI
Non	300	Medium	0.98	0.84	0.78	0.23	0.08	0.03
		High	1.00	1.00	1.00	0.34	0.11	0.02
	500	Medium	0.98	0.96	0.78	0.14	0.04	0.01
		High	1.00	1.00	1.00	0.17	0.04	0.008
	1000	Medium	1.00	1.00	0.91	0.10	0.01	0.00
		High	1.00	1.00	1.00	0.10	0.02	0.00
Impact	300	Medium	0.89	0.78	0.64	0.26	0.07	0.06
		High	1.00	1.00	0.94	0.25	0.09	0.03
	500	Medium	0.97	0.96	0.73	0.16	0.04	0.005
		High	1.00	1.00	1.00	0.25	0.11	0.009
	1000	Medium	1.00	1.00	0.89	0.11	0.02	0.00
		High	1.00	1.00	1.00	0.12	0.03	0.00

DISCUSSION

Tests and questionnaires to measure attitudes or personality usually use Likert or ordered-categorical response format, but they are analyzed using models that assume the continuous character of those variables. There are very important studies describing those models, but the literature does not have a lot of work about using them to test measurement invariance. The aim of this work was to assess the power and Type I error rates in detecting ME using one factor linear model for ordered categorical items. There are other works that analyze the effectiveness of the continuous linear models in detecting DIF (Meade and Lautenschlager, 2004b; Oort, 1998;

Raju, Laffite, and Byrne, 2002; Stark, Chernyshenko, and Drasgow, 2006), but the literature has not paid much attention to the use of the ordered categorical models (French and Finch, 2006). However, several works warned about the threats of using the continuous model in categorical or Likert type data (Millsap and Yun-Tein, 2004; Lubke and Muthén, 2004). The continuous and ordered models are not equivalent. The latter includes thresholds, which are inherent to Likert type data, but using a continuous linear model is not possible to assess its equivalence between groups. Theoretically, the ordered categorical models should be preferred, but this is not what is found in the applied context. The researchers used continuous linear models to analyze Likert type data although they were violating the condition of multivariate normality.

We wanted to add one point to the literature about ordered-categorical models and to show the utility of this model in assessing measurement invariance. We simulated several conditions, varying the sample size, type of DIF, amount of DIF, and presence of impact. We used a strategy named the “constrained baseline model” (Stark, Chernyshenko, and Drasgow, 2006), and we used three different indexes for evaluating power and Type I error rates. Under this strategy and evaluated conditions, the power of DIF detection using chi-square test was very good, but these criteria did not control the Type I error. The results suggested the need to use more restrictive criteria in flagging items. The use of Bonferroni correction showed very good control of the false positives and its power was close to 1.00 in all conditions. Only when the sample size was small (N=300) and the amount of DIF medium (0.25; uniform and non uniform), did the power not reach a value of 0.90. The use of the CFI difference index also reduced the false positives, even more than Bonferroni correction. The power rates of this index were slightly less than with the Bonferroni corrections, especially in presence of impact and medium amount of DIF. When the amount of DIF is big, the power of the CFI is close to 1.00. Although this index is slightly more conservative than Bonferroni correction, its characteristics turns it into a very good practical criterion for the detection of DIF.

The results reported in this study confirm the adequacy of the common factor model for ordered categorical data in assessing DIF. Those results are not concordant with the findings from French and Finch’s (2006). They worked with multidimensional dichotomous data, different sample sizes and percentages of DIF bigger than we defined in this study. The number of indicators for factors was also different; they defined two conditions with 3 and 6 indicators whereas we worked with 9, which is a number extracted from real data. The works were different in a lot of aspects so more studies would be need in order to evaluate the effect of each of the factors on the final results.

Of course, our work had several limitations. Maybe one of the most important was the definition of only one DIF item in the test. Although most of the works assessing DIF in polytomous items use this design (French and Miller, 1996; Kristjanssson, Aylesworth, Dowell, and Zumbo, 2005; Spray and Miller, 1994; Zwick, Donogue and Grima, 1993; Zwick, Thayer and Mazzeo, 1997), it is difficult to assume than only one item presents DIF in practical situations. The detection of DIF when more than one item is functioning differently in two groups adds another important aspect to the problem of assessing DIF using the difference of fit between nested models, which is the definition of the baseline model. The strategy that we used consisted of comparing nested models: one baseline model against one model with one item parameters free. This strategy was evaluated for all items that were components of the test (except for the referent item). This way of assessing DIF is followed by one of the most popular IRT approaches, the Likelihood Ratio test (LR). The basis of the LR method is similar to the procedure described in this paper (Cohen, Kim, and Baker, 1993; Thissen, 1991; Thissen, Steinberg, and Wainer, 1998). The LR compares nested models, wherein parameters are fixed or free. The baseline model is defined by constraining the parameters for all items. Different models are formed by freeing the parameters for the evaluated item, one at time. Then, the variation in G2 is evaluated. Recent works have shown that it would be more statistically correct (Maydeu-Olivares and Cai, 2006) to use the baseline free model. This idea would be more important in the conditions where more than one item had DIF.

In this work we used WLS estimation which performs well with dichotomous and ordinal variables (Jöreskog, 1994). However robust weighted least squares (RWLS) estimation is being recommended by some authors (Muthén, du Toit and Spisic, 1997; Finney and DiStenfano, 2006; Beauducel and Herzberg, 2006; Flora and Curran, 2004) . WLS requires the weight matrix to be positive definite because the weight matrix is inverted as part of the estimation procedure, and when small samples are used WLS can lead to estimation problems. RWLS use the diagonal matrix and the estimates are reasonably stable even with small samples (N=100). In this simulation we didn’t have converge problems, but would be interesting to evaluate the results using RWLS.

Also, it would be convenient to follow in this line of work and extend the study in different ways. It would be very interesting to assess the similarities and dissimilarities between the linear continuous and linear ordered models in the study of invariance. Even using ordered linear models, we found two different approaches depending on which software was used. Two of the most implemented software packages, LISREL (Jóreskog and Sörbom, 1993) and Mplus (Muthén and Muthén, 2004), used different baseline models and different parameterization in the definition of the invariance for ordered categorical data (Millsap and Yun-Tein, 2004). Another issue that remains important is the comparison between these models and the item response theory models. The work that compares those two approaches (Meade and Lautenschlager, 2004a, 2004c; Raju, Laffite, and Byrne, 2002; Stark, Chernyshenko, and Drasgow, 2006) used the continuous model. It would be interesting to have more information regarding those two models (linear/non linear) which are designed to model the ordered categorical responses. In summary, the results reported were good and promising under the simulated condition, and so we can say that the common factor model for ordered responses can be successfully used to analyze invariance between groups.

RESUMEN

Equivalencia métrica en datos categóricos ordenados. La invarianza factorial estudia de la equivalencia métrica en el marco del modelo lineal del factor común por medio de la comparación de los parámetros del modelo de medida en los grupos de interés. Sin embargo cuando se utilizan ítems dicotómicos, Likert o categorías de respuestas ordenadas se viola la asunción referida al carácter continuo de las variables. Aunque existen modelos explícitos para este tipo de datos son muy escasos los trabajos que analizan su potencia y error Tipo I en el estudio de la invarianza factorial. Por medio de simulación Montecarlo este trabajo analiza la potencia y error tipo I asociados a la detección de la invarianza factorial en un diseño que manipula cuatro factores; tamaño de la muestra (300, 500 y 1000), tipo de DIF (umbrales, pesos), cantidad de DIF (0,25, 0,40), y presencia de impacto. Los parámetros de generación de datos provienen de una escala unifactorial compuesta por 9 indicadores con 3 categorías de respuesta ordenada. La presencia/ausencia de invarianza se evaluó utilizando tres criterios : a) significación de la diferencia entre valores chi-cuadrado de modelos anidados, b) la significación de la diferencia entre valores chi-cuadrado de modelos anidados aplicando la corrección Bonferroni, y c) la diferencia entre los valores del Índice Comparativo de Ajuste (CFI) entre modelos anidados. Los resultados mostraron un buena potencia y control de falsos positivos asociados a la diferencia entre CFIs y a la corrección Bonferroni.

REFERENCES

Beauducel, A., and Herzberg, P.Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, 13, 186-203.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bock, R. D., and Aitkin, M. (1981). Marginal maximum likelihood estimation or item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443-459.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, Wiley.

Cattell, H. B. (1989). *The 16PF, Personality in depth*. Champain, IL, Institute for Personality and Ability Testing, Inc.

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40, 5-32.

Cohen, A. S., Kim, S. H., and Baker, F. D. (1993). Detection of differenttial item functioning in the graded response model. *Applied Psychological Measurement, 17*(4), 335-350.

Elosua, P., and López, A. (2008). Adapting the AF5 autoconcept scale to Basque: Validity evidences. *Individual Differences Research, 6(1),57-70.*

Elosua, P., and Wells, C. (July, 2008). A Comparison of MACS and the IRT Likelihood Ratio Test for Identifying DIF. *Paper presented at III European Congress of Methodology*, Oviedo.

Finch, W.H., and French, B.F. (2007). Detection of Crossing Differential Item Functioning. A Comparison of four Methods. *Educational and Psychological Measurement,* 67(4), 565-582.

Finney, S.J., and DiStenfano, C. (2006). Non-normal and categorical data in structural equation modelling. In G.R. Hancock & R. O. Mueller (Eds*.), Structural Equation Modeling: A second course* (pp. 269-314). Information Age Publishing Inc.

Flora, D.B., and Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466-491.

French, B.F., and Finch, W.H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, 13(3), 378-402.

French, A. W., and Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3), 315-332.

Ihaka, R., and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5, 299–314.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36, 409-426.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, 24, 387-404.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika 59,* 381-389.

Jöreskog, K. G., and Sörbom, D. (1993) *LISREL 8: User's guide*. Chicago: Scientific Software International.

Kristjansson, E., Aylesworth, R., McDowell, I., and Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935-953.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ, LEA.

Lubke, G. H., and Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling, 11*(4), 514-534.

Maydeu-Olivares, A., and Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*, 55-64.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Meade, A. W., and Lautenschlager, G. J. (2004a). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361-388.

Meade, A. W., and Lautenschlager, G. J. (2004b). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11*, 60-72.

Meade, A. W., and Lautenschlager, G. J. (2004c). Same question, different answer: CFA and two IRT approaches to measurement invariance. *19th Annual Conference of the Society for Industrial and Organizational Psychology*, Chicago.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika,* 29, 177-185.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 521-543.

Millsap, R.E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares and J. J. McArdle (Eds.), *Contemporary psychometrics* (pp.153-172). Mahwah, NJ: Lawrence Erlbaum

Millsap, R. E., and Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11*, 3-31.

Musitu, G., García, F., and Gutiérrez, M. (1997) *AFA. Autoconcepto Forma-A.* Madrid: TEA

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, 49, 115-132.

Muthén B. O., du Toit S. H. C., and Spisic D.*(*1997*). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes.* Unpublished manuscript*,* University of California*,* Los Angeles*.*

Muthén, B. O., and Kaplan, D. (1985). A comparison of some methodologies for the factor análisis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, 38, 171-189.

Muthén, L. K., and Muthén, B. O. (2004). *Mplus users's guide*. Los Angeles: Muthén and Muthén.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107-124.

Raju, N. S., Laffite, L. J., and Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517-529.

Schuerger, J. M. (2001). *16PF-APQ Manual*. Champaign, IL, Institute for Personality and Ability Testing.

Sörbom, D. (1974) A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology,* 27, 229-239.

Stark, S., Chernyshenko, O. S., and Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Spray, J. A., and Miller, T. (1994*). Identifying nonuniform DIF in polytomously scored test items.* (American College Testing Research Report Series 94-1)Iowa City, IA: American College Testing Program.

Thissen, D. (1991). *MULTILOG user's guide. Item analysis and scoring with multiple category response models (Version 6).* Mooresville, IN, Scientific Software.

Thissen, D., Steinberg, L., and Wainer, H. (1988). Use of item respones theory in the study of group differences in trace lines. In H. Wainer and H. I. Braun (Eds.) *Test validity*. (pp. 147-169) Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R., Donogue, J. R., and Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30*(3), 233-251.

Zwick, R., Thayer, D. T., and Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10*(4), 321-344.

(Manuscript received: 5 November 2009; accepted: 1 October 2010)